Quantitative Trait Linkage Analysis for Different Human Family Types

Ushma Galal and Lize van der Merwe
South African Medical Research Council
Cape Town, SOUTH AFRICA
Corresponding author: Ushma Galal, email: ushma_galal@mrc.ac.za

Abstract

Data from extended families is often useful for assessing the genetic effect on a potentially heritable numerical (quantitative) response (trait). However, accounting for relatedness complicates a statistical analysis because it requires specialised mixed-effects modelling. In addition, families in a study group may differ in size, further complicating the modelling. The statistical theory underlying the genetic linkage analysis model is not easily understood and has not been comprehensively documented, creating confusion about why this data cannot be modelled using standard statistical software. We introduce the necessary genetic terminology and concepts then explain the genetic linkage analysis mixed-effects model for progressively more complex family types.

Key Words: Extended families, genetic association, variance components

1. Introduction

Human genetic material consists of 23 pairs of chromosomes. We inherit one member of each pair from our mothers and the other from our fathers. In this study, we consider only the 22 autosomal pairs, for which each paired member is identical to its counterpart; that is, they contain the same genes (pieces of genetic material that may code for a biological function) in the same positions, referred to as gene *loci* (singular: *locus*). The genetic material contained by the paired loci may differ across the population so each version is called an *allele* of that locus. An individual's *genotype* (genetic status at a locus) is often denoted by paired alleles, for example $1/2, 1/1, 2/2$. If the two alleles differ, then it is normally assumed that only one of them is correlated to the trait of interest and is inherited together with the trait, from the same parent. We are essentially searching for this allele. The paired alleles found at every gene locus in an individual are considered to be independent of each other unless the individual's parents are related. Thus, the allele pairs inside each family member are independent, but the people involved are not. In reality, we do not inherit our parental genomes intact. Pieces of our maternal chromosomes break off and swap with the corresponding pieces of our paternal chromosomes, resulting in siblings inheriting different combinations of their parents' genomes.

Linkage occurs when two *markers* (genetic factors with known genomic locations) are inherited together in a way that suggests they are joined or linked to each other. It is based on the idea that the markers are close enough to each other on the genome so that breakage is not likely to occur between them, resulting in them being inherited intact from the same parent.

Linkage analysis is the method used to link a trait to a particular marker. It is based on the fact that in a family, a marker and trait-affecting locus are inherited together more often than is expected by chance. Using this, patterns of inheritance can be traced in families because relatives with highly correlated heritable trait values should also be more highly correlated genetically. These

correlation patterns together with our markers help us to narrow down regions on the genome in which trait-affecting alleles may lie. If an allele at a specific marker is responsible for increasing (or decreasing) values of a trait, then in the presence of linkage, family pairs who have high (or low) trait values are expected to share the trait-affecting allele.

A statistical analysis should account for the fact that relatives are related and are genetically similar as this will account for some of the variation observed in the data. Family relatedness is quantified through a per-family random effect and corresponding *kinship* coefficient matrix. Genetic similarities at a particular locus are captured though a locus-specific random effect and corresponding coefficient matrix. The linkage model we explain underlies, among others, the QTDT (Quantitative Transmission Disequilibrium Tests) software (Abecasis et al., 2000a,b) and the *lmekin* function found in the Therneau (2012) package of the R software (R Development Core Team, 2011).
Refer to Elston (2000) and Burton et al. (2005) for more detailed descriptions of the linkage concept and genetic terminology introduced here.

2. Kinship coefficients and Identity by descent
$\varphi_{ijk} = (\frac{1}{2})^{R+1}$ denotes the *kinship coefficient* between members j and k of family i, where R denotes the degree of relationship between them and $(\frac{1}{2})^R$ is the expected proportion of shared alleles at any locus (Thomas, 2004).

The degree of genetic similarity between family pairs at a specific genetic locus is quantified through a measure based on allelic sharing, defined as follows:
The probability that two family members share 0, 1 or 2 alleles from the same source, at a specific locus, defines their genetic relationship at that locus. Two alleles at a specific locus, each one from a different person, are said to be *identical-by-descent* (IBD) if they come from a common ancestor. We let $\pi_{ijk}(a)$ denote the probability that individuals j and k in family i share a alleles IBD at a particular genetic locus. Thus the expected proportion of alleles they share IBD can be denoted by $\bar{\pi}_{ijk} = \frac{1}{2}\sum_{a=0}^{2} a.\pi_{ijk}(a) = \frac{1}{2}[1.\pi_{ijk}(1) + 2.\pi_{ijk}(2)]$.

Suppose that genotype information is available for a specific marker for the extended family in Figure 1, where Dad and Mom are the parents of Sue and Jane; Ryan is married to Jane and Ally is their daughter. We assume that Ryan is only related to the rest of the family through his marriage and his genotype is unknown (0/0). Table 1 shows the kinship coefficients and inferred IBD sharing of pairs of members from this family.

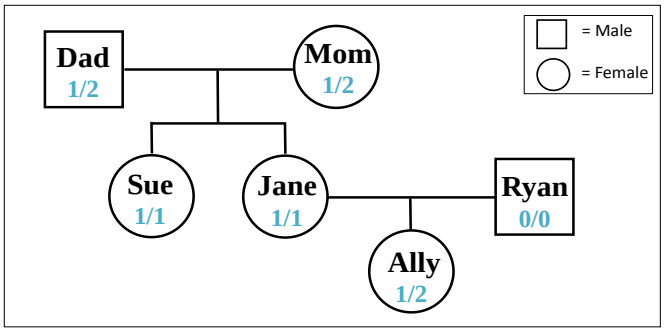

Figure 1: Genotypes for a hypothetical extended family

Table 1: IBD probabilities and kinship coefficients for the hypothetical extended family in Figure 1

			IBD probabilities				
Relative type	**Relative pair**	**R**	$\pi(0)$	$\pi(1)$	$\pi(2)$	$\bar{\pi}_{ijk}$	2φ
Parent–offspring	Dad/Mom–Sue/Jane	1	0	1	0	1/2	1/2
	Jane/Ryan–Ally	1	0	1	0	1/2	1/2
Sib–pairs	Sue–Jane	1	0	0	1	1	1/2
Grandparent–grandchild	Dad–Ally	2	1	0	0	0	1/4
	Mom–Ally	2	0	1	0	1/2	1/4
Aunt–niece	Sue–Ally	2	0	1	0	1/2	1/4
Unrelated individuals	Dad–Mom	∞	1	0	0	0	0
	Dad/Mom/Jane/Sue–Ryan	∞	1	0	0	0	0

In Table 1, Dad and Ally share 0 alleles IBD for the marker, but Mom and Ally share 1 allele IBD, even though both pairs are grandparent-grandchild relationships. The two pairs have different IBD sharing at this locus, but we still capture their relationship through the kinship coefficient, which is the same for both pairs. In addition, although Ryan's genotypes are unknown, we can still calculate his IBD sharing with the rest of the family because of his daughter.

When some genotypes are not available or informative, IBD sharing cannot always be inferred. In such cases, theoretical or prior IBD probabilities have to be used. These are the corresponding kinship coefficients since they are based on the same information.

3. Variance components-based linkage analysis

Linkage analysis uses the relationships between relative-pairs and allele-sharing at a specific marker to locate trait-affecting loci. It is based on the alternative hypothesis that, for markers neighbouring a candidate trait locus, genotypic similarities are positively correlated with trait similarities for relative pairs that share alleles IBD (Almasy & Blangero, 2010).

Let the variance components be denoted as follows: σ_a^2 is the locus-specific variance, σ_g^2 is the hereditary variance (due to all genetic factors other than the locus-specific one) and σ_e^2 is the environmental (non-genetic) variance consisting of non-heritable shared-environment ("cluster") variance, unshared or unique environment variance and residual variance due to random errors. The kinship coefficients extract the hereditary variance from the total variance while the IBD sharing extracts the locus-specific variance. This allows us to separate out the effects of the variance attributable to family relatedness from that of the alleles inherited at a specific locus.

To model linkage, consider a set of r families with n_i members such that $N = \sum_i^r n_i$ gives the total number of individuals in the study group. Let $\mathbf{y_{ij}}$ represent the random quantitative trait value for individual j, from family i. Let $\underline{\mathbf{y}}_\mathbf{i}(n_i \times 1) = (\mathbf{y_{i1}}, \ldots, \mathbf{y_{in_i}})^T$ be the random vector of trait values for family i and let the overall vector of trait values be $\underline{\mathbf{y}}(N \times 1) = (\underline{\mathbf{y}}_\mathbf{1}, \ldots, \underline{\mathbf{y}}_\mathbf{r})^T$. Let $\mathbf{g_i} \sim$ i.i.d $\mathcal{N}(0, \sigma_g^2)$ denote the hereditary random effect for family i, and let $\mathbf{e_{ij}} \sim$ i.i.d $\mathcal{N}(0, \sigma_e^2)$ be the environmental effect for individual j from family i. Now let $\mathbf{a_i} \sim$ i.i.d $\mathcal{N}(0, \sigma_a^2)$ be the random locus-specific additive allelic effect for family i. It measures the impact of the additive allelic effect at a specific marker

locus. Assume the three variance components, σ_a^2, σ_g^2 and σ_e^2 are mutually independent so that the total variance is their sum.
For the j^{th} individual in the i^{th} family, a model for this data is

$$\mathbf{y_{ij}} = \mu + \underline{z}_{a_ij}^T \underline{\mathbf{a}}_\mathbf{i} + \underline{z}_{g_ij}^T \underline{\mathbf{g}}_\mathbf{i} + \mathbf{e_{ij}}, \text{ for } i = 1, \ldots, r;\ j = 1, \ldots, n_i, \tag{1}$$

where the superscript 'T' indicates a transpose, μ is the overall mean trait value; $\underline{z}_{g_ij}^T(1 \times n_i) = \{z_{g_ijk}\}$ is the vector of regression coefficients for individual j, corresponding to $\mathbf{g_i}$; and the vector of regression coefficients for individual j, corresponding to $\mathbf{a_i}$, is given by $\underline{z}_{a_ij}^T(1 \times n_i) = \{z_{a_ijk}\}$.
Extending (1) for the i^{th} family gives

$$\underline{\mathbf{y}}_\mathbf{i} = \underline{1}_{n_i}\mu + \mathcal{Z}_{a_i}\underline{\mathbf{a}}_\mathbf{i} + \mathcal{Z}_{g_i}\underline{\mathbf{g}}_\mathbf{i} + \underline{\mathbf{e}}_\mathbf{i}, \text{ for } i = 1, \ldots, r, \tag{2}$$

where $\underline{1}_{n_i}$ is the $(n_i \times 1)$ vector of 1's; $\mathcal{Z}_{g_i}(n_i \times n_i) = \{z_{g_ijk}\}$ and $\mathcal{Z}_{a_i}(n_i \times n_i) = \{z_{a_ijk}\}$ are the matrices of regression coefficients for $\mathbf{g_i}$ and $\mathbf{a_i}$, respectively.
By the assumption of mutually independent random effects, $\underline{\mathbf{y}}_\mathbf{i} \sim$ i.i.d $\mathcal{N}(\underline{1}_{n_i}\mu, \mathbf{\Omega_i})$, where $\mathbf{\Omega_i}(n_i \times n_i) = \sigma_a^2 \mathcal{Z}_{a_i}\mathcal{Z}_{a_i}^T + \sigma_g^2 \mathcal{Z}_{g_i}\mathcal{Z}_{g_i}^T + \sigma_e^2 \mathcal{I}_{\mathbf{n_i}}$ and $\mathcal{I}_{\mathbf{n_i}}$ is the $n_i \times n_i$ identity matrix.

For Model (2), the coefficients of the covariance matrices for the hereditary and locus-specific random effects are the kinship coefficients and IBD sharing for each pair of family members, respectively, where

$$\mathcal{Z}_{g_i}\mathcal{Z}_{g_i}^T = \begin{pmatrix} 1 & 2\varphi_{i12} & \cdots & 2\varphi_{i1n_i} \\ 2\varphi_{i21} & 1 & \cdots & 2\varphi_{i2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ 2\varphi_{in_i1} & 2\varphi_{in_i2} & \cdots & 1 \end{pmatrix}, \mathcal{Z}_{a_i}\mathcal{Z}_{a_i}^T = \begin{pmatrix} 1 & \bar{\pi}_{i12} & \cdots & \bar{\pi}_{i1n_i} \\ \bar{\pi}_{i21} & 1 & \cdots & \bar{\pi}_{i2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{\pi}_{in_i1} & \bar{\pi}_{in_i2} & \cdots & 1 \end{pmatrix}.$$

Thus, the covariance between individuals j and k in family i is

$$\Omega_{ijk} = cov(\mathbf{y_{ij}}, \mathbf{y_{ik}}) = \begin{cases} \sigma_a^2 + \sigma_g^2 + \sigma_e^2 & \text{if } j = k \\ \bar{\pi}_{ijk}\sigma_a^2 + 2\varphi_{ijk}\sigma_g^2 & \text{if } j \neq k. \end{cases}$$

Using the family in Figure 1 and the information in Table 1 we illustrate the covariance matrices corresponding to the three variance components, for different family types. For all the family-types, the covariance matrices corresponding to the random environmental effect are identity matrices.
Sib-pair (Sue and Jane):

$$\mathbf{\Omega_i}(2 \times 2) = \sigma_a^2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \sigma_g^2 \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix} + \sigma_e^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{3}$$

Parent-offspring trio (Jane, Ryan and Ally):

$$\mathbf{\Omega_i}(3 \times 3) = \sigma_a^2 \begin{pmatrix} 1 & 0 & \frac{1}{2} \\ 0 & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix} + \sigma_g^2 \begin{pmatrix} 1 & 0 & \frac{1}{2} \\ 0 & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix} + \sigma_e^2 \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{4}$$

Nuclear family (Dad, Mom, Sue and Jane):

$$\mathbf{\Omega_i}(4 \times 4) = \sigma_a^2 \begin{pmatrix} 1 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 & 1 \\ \frac{1}{2} & \frac{1}{2} & 1 & 1 \end{pmatrix} + \sigma_g^2 \begin{pmatrix} 1 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix} + \sigma_e^2 \mathcal{I}_\mathbf{4} \tag{5}$$

Extended family (Dad, Mom, Sue, Jane, Ryan and Ally):

$$\begin{aligned}\mathbf{\Omega_i}(n_i \times n_i) &= \sigma_a^2 \begin{pmatrix} 1 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 & 1 & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 & 1 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 1 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix} + \sigma_g^2 \begin{pmatrix} 1 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{4} \\ 0 & 1 & \frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & 1 & \frac{1}{2} & 0 & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 1 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix} \\ &+ \sigma_e^2 \mathcal{I}_{\mathbf{n_i}}. \end{aligned} \tag{6}$$

In Equation (3), the coefficient matrices for both σ_a^2 and σ_g^2 have a compound symmetric structure, and the coefficient matrix of the former is based on Sue and Jane's IBD sharing at the locus. In practice, we need to know the IBD sharing of every sib-pair in a study group before we model the data so that the coefficients of the covariance matrices can be specified accordingly.

For both the trio and nuclear family in Equations (4) and (5) respectively, the zeros in the coefficient matrices of σ_a^2 and σ_g^2 represent the relationship between the unrelated parents. These zeros cause the compound symmetric structure to break down. As with the sib-pairs, we need to know the genotype-based IBD sharing beforehand for each family in our study group and we need to find a way to incorporate this information into our model. For trios, the IBD matrix for parents and one child is always the same, because children must share one allele IBD with each parent, with a probability of 1.

The coefficient matrices for σ_a^2 and σ_g^2 for the extended family in Equation (6) do not have recognisable structures. In addition, because extended families are not generally the same size, the dimensions of the covariance matrices will differ accordingly, preventing us from modelling them explicitly.

For a sample of r families of arbitrary size, let $\mathbf{\Omega_i}$ represent the per-family covariance matrices and let $\mathcal{O}_{\mathbf{is}}$ denote zero-matrices of appropriate dimension. The covariance matrix, $\mathbf{\Omega}$, for $\underline{\mathbf{y}}(N \times 1)$ is a large block-diagonal matrix where each block is a family-specific covariance matrix:

$$\mathbf{\Omega} = \begin{pmatrix} \mathbf{\Omega_1} & \mathcal{O}_{\mathbf{12}} & \cdots & \mathcal{O}_{\mathbf{1r}} \\ \mathcal{O}_{\mathbf{21}} & \mathbf{\Omega_2} & \cdots & \mathcal{O}_{\mathbf{2r}} \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{O}_{\mathbf{r1}} & \mathcal{O}_{\mathbf{r2}} & \cdots & \mathbf{\Omega_r} \end{pmatrix}.$$

In a linkage model, we estimate only the three variance-components σ_a^2, σ_g^2 and σ_e^2, and test σ_a^2. If it is not significant, then under the null hypothesis, there is no evidence for linkage in the study group. The linkage model can be run in software programs such as QTDT and *lmekin* in R, both of which can also calculate kinship matrices for various family types. Estimating IBD sharing is more complicated and must be carried out elsewhere before being imported into the relevant package.

4. Conclusion

Linkage analysis of quantitative trait data requires specialised mixed-effects model software because the modelled correlations are potentially different for each pair of family members. This is due to the correlations being dependent on the IBD sharing and kinship coefficients between each family pair. Furthermore, only three

specific variance components are estimated; namely the locus-specific, hereditary and environmental variances, by using the IBD sharing and kinship coefficients to extract the locus-specific and hereditary variances, respectively, from the data. In addition, the families observed are not usually the same size and extended families are preferred as they are more informative for linkage than smaller family units (Blangero et al., 2001). These factors combined have resulted in the development of specialised software for this type of analysis.

The linkage model shown here can be extended to include other variance components that represent, for example, shared environmental effects. Refer to Almasy & Blangero (2010) for more information. Lastly, in practice we usually adjust for covariates and estimate their effects; however, since our focus was on variance components-based linkage analysis, covariates were not considered in this study.

References

Abecasis, G. R., Cardon, L. R., & Cookson, W. O. C. (2000a). A general test of association for quantitative traits in nuclear families. *American Journal of Human Genetics*, 66, 279–292.

Abecasis, G. R., Cookson, W. O. C., & Cardon, L. R. (2000b). Pedigree tests of transmission disequilibrium. *European Journal of Human Genetics*, 8, 545–551.

Almasy, L. & Blangero, J. (2010). Variance components methods for analysis of complex phenotypes. *Cold Spring Harbor Protocols*, 5, pdb.top77.

Blangero, J., Williams, J., & Almasy, L. (2001). Variance component methods for detecting complex trait loci. In *Advances in Genetics*.

Burton, P. R., Tobin, M. D., & Hopper, J. L. (2005). Key concepts in genetic epidemiology. *Lancet*, 366, 941–951.

Elston, R. C. (2000). Introduction and overview. *Statistical Methods in Medical Research*, 9, 527–541.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Therneau, T. (2012). *coxme: Mixed Effects Cox Models.* R package version 2.2-3.

Thomas, D. C. (2004). *Statistical methods in genetic epidemiology.* New York: Oxford University Press Inc.